Serie Apuntes de Clase ΩBΓ

N° 11
Enero del 2015

INTRODUCCIÓN AL MANEJO DE BASE DE DATOS EN STATA 14

Rafael Bustamante Romaní

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, DECANA DE AMÉRICA

FACULTAD DE CIENCIAS ECONÓMICAS

La **Serie Apuntes de Clase Omega Beta Gamma** tiene por objetivo difundir los materiales de enseñanza generados por los docentes que tienen a su cargo el desarrollo de las asignaturas que forman parte de los Planes de Estudios de las Escuelas Académico-Profesionales de la Facultad de Ciencias Económicas de la Universidad Nacional Mayor de San Marcos. Estos documentos buscan proporcionar a los estudiantes la explicación de algunos temas específicos que son abordados en su formación universitaria.

Encargados de la serie:

Bustamante Romaní, Rafael.
rbustamanter@unmsm.edu.pe

Cisneros García, Juan Manuel.
jcisnerosg@unmsm.edu.pe

Facultad de Ciencias Económicas.
Universidad Nacional Mayor de San Marcos.
Calle Germán Amézaga N° 375.
Ciudad Universitaria, Lima 1. Perú.

La **Serie Apuntes de Clase ΩBΓ** es promovida y desarrollada por un colectivo de docentes del Departamento de Economía de la Universidad Nacional Mayor de San Marcos.

Introducción al uso de base de datos en STATA 14

Rafael Bustamante[◊]

Resumen

Este trabajo busca explicar los procedimientos básicos en el uso del paquete Stata 14. Exploración básica del paquete así como estudio de algunos estadísticos descriptivos, gráficos, manejo de base de datos. Asimismo se da una explicación básica de los comandos más comunes en uso. El objetivo final de este documento es que el lector aprenda a interactuar con el software.

Palabras Claves: Estadísticos Descriptivos, Stata.

Clasificación JEL: C2, C25

[◊] Doctorado en Economía con mención en los Recursos Naturales (c), Universidad Nacional Autónoma de México. MBA Gerencial, CENTRUM Pontificia Universidad Católica del Perú. Maestría en Economía con mención en Finanzas, Universidad Nacional Mayor de San Marcos. B. Sc. Economía, UNMSM. Profesor Auxiliar del Departamento de Economía de la UNMSM. Investigador asociado al Instituto de Investigaciones FCE - UNMSM. Contacto: rbustamanter@unmsm.edu.pe

I. EL ENTORNO DE STATA 14

Stata es un programa estadístico para investigadores de diferentes disciplinas, como bioestadísticos, investigadores sociales y económicos. Los diferentes tipos de análisis integrados a Stata están documentados y soportados teóricamente por numerosos documentos, publicaciones y revistas. Además los manuales de Stata reúnen en 21 volúmenes con ejemplos estadísticos, explicaciones teóricas, métodos, fórmulas y documentos de referencia (ver www.stata.com/manuals/). Al tratarse de un programa en ambiente Windows, su interface es similar a la de todos los programas bajo este ambiente. **Stata** está disponible en 4 tipos de versión (Rojas & Gordillo, 2012).
A continuación presentamos el siguiente cuadro resumen del entorno de Stata 14.

Figura Nº1

Small Stata	Versión estudiantil de **Stata**
Intercooled Stata	Versión estándar de **Stata**
Stata/SE	Versión especial de **Stata** para manejo de bases de datos grandes.
Stata/MP	Versión especial de **Stata** diseñada para trabajar en equipos con más de un procesador o núcleo (2 a 32 procesadores)

El despliegue de Stata presenta cuatro ventanas diferentes:

Figura Nº2

Review	Aquí aparecen los comandos que han sido utilizados durante la sesión. Solo los resultados más recientes son visibles en esta pantalla. Generalmente se usa cuando se activan las bitácoras.
Command	Sirve para utilizar Stata de forma interactiva, es decir se emplea para crear las líneas de comandos y llevar a cabo las aplicaciones disponibles en el software.

Variables	Nos informa sobre las variables que están disponibles en nuestra base de datos para realizar las diversas aplicaciones
Results	Esta nos permite visualizar los resultados (outputs) de los estadísticos que pedimos calcular o de los modelos que solicitamos estimar.

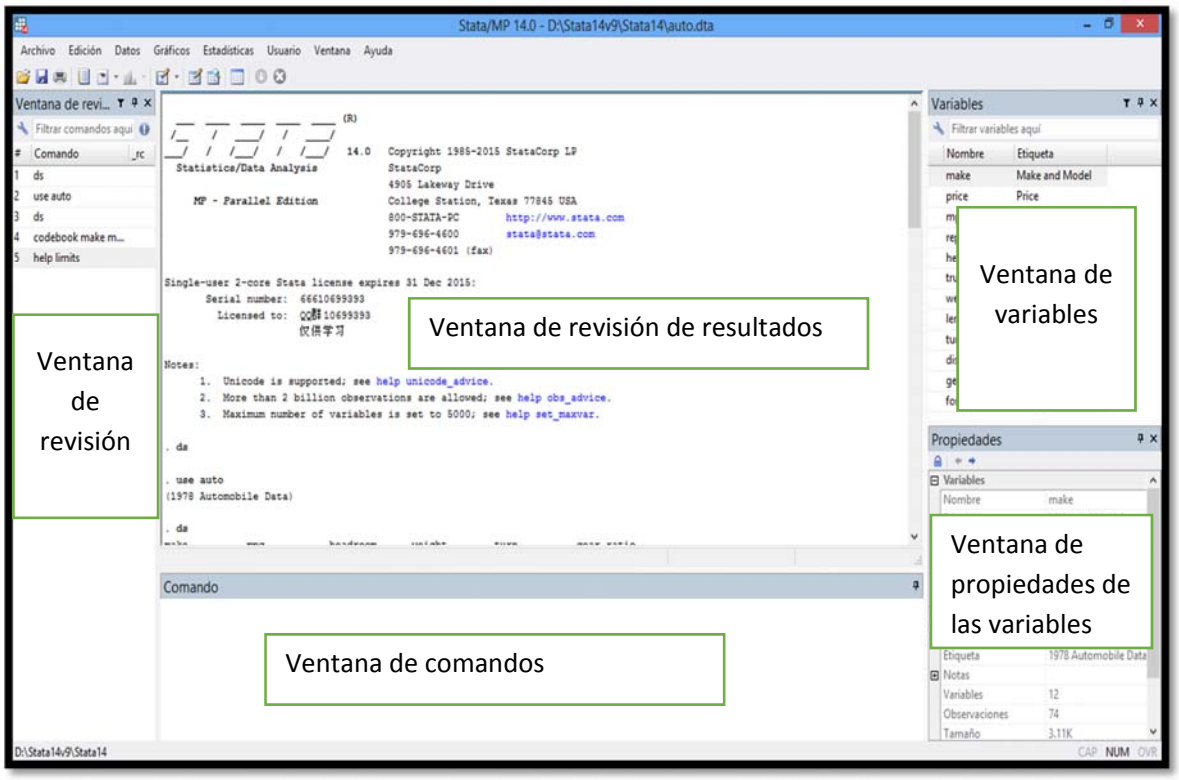
***Stata Editor*:** Permite navegar y modificar los datos como si fuese una hoja de cálculo Excel.
***Stata Viewer*:** Permite acceder a información en línea y también a la ayuda del programa.
Stata Graphs: Presenta el último gráfico realizado.

***Stata Do-file Editor*:** es una ventana separada en Windows y funciona como un editor de textos que permite ejecutar una lista de comandos.

Stata Browser*[1]*: Permite visualizar los datos, mas no modificarlos.

[1] Si alguna ventana está cerrada podemos abrirla desde el menú Windows.

Figura Nº 3



Ventana de Variables: Muestra el listado de variables de la base de datos activa.
Ventana de Comandos: Se describen y almacenan las líneas de comandos, si se desea recuperar un comando previo puede utilizar las teclas RePág o AvPág y podrá autocompletar el nombre de la variable utilizando la tecla TAB.
Ventana de Resultados: Permite visualizar la sintaxis, y los resultados de los procedimientos ejecutados por el usuario. Aquí encontrará el logo de **Stata**, indicando la versión y el tipo de licencia y el número máximo de variables a importar. Una de las características de ésta ventana es que por medio de colores el programa informa si un comando ha sido correctamente ejecutado, si aparece en color negro no hubo problema en la realización, rojo indicar error y el azul es un hipervínculo al menú de ayuda.

Ventana de Revisión: Bitácora que permite llevar un completo registro de todos los procedimientos ejecutados durante una sesión de **Stata** ya sea que se ejecutaron por el ambiente GUI, por la ventana de comandos o por un editor .do. Una de las

propiedades de la ventana *Review* es que si se desea repetir un comando simplemente debe hacer doble clic sobre el comando deseado y **Stata** lo ejecutará de nuevo.

Ventana de Propiedades: Presenta la información de cada variable, como nombre, tipo de variable, formato, las notas de la base de datos (puede usar el comando notes para verlas en la ventana de resultados), entre otras características.

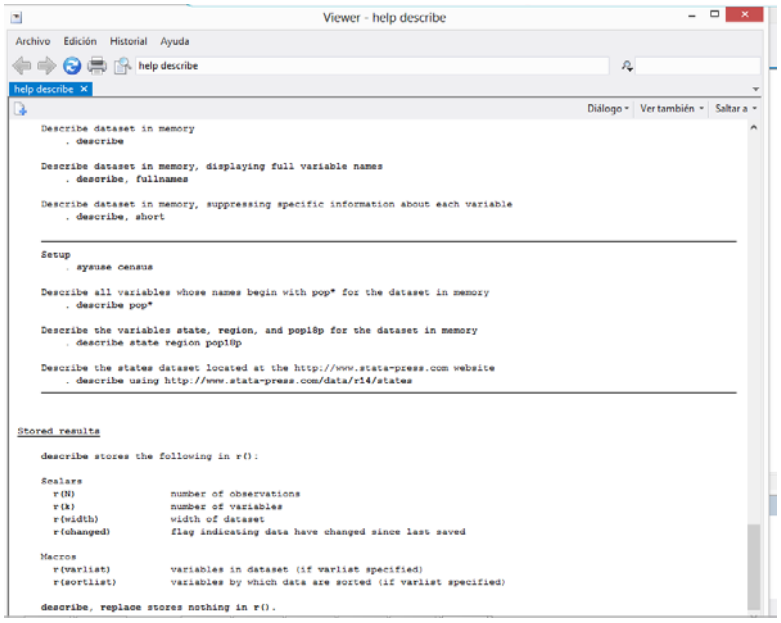
II. EL MENÚ DE AYUDA

Stata ha incorporado a partir de la versión 11 un conjunto de nuevas opciones en el menú de ayuda para facilitarle al usuario la mejor forma de entender cómo funciona el programa. Una de las novedades principales es que se ha agregado la opción de tener disponibles los manuales de Stata en formato PDF. Para acceder a los manuales de ayuda debe seguir la ruta Help ➔ PDF Documentación. El menú de ayuda de Stata le permite:

- Ver el índice de contenidos del programa.
- Buscar información sobre algún tema, la rutina que permite ejecutarla en Stata, o el sitio desde donde es posible descargar la macro para alimentar el programa.
- Obtener ayuda sobre algún comando de Stata.
- Listar y descargar las últimas actualizaciones del programa.
- Instalar programas de Stata escritas por otros usuarios, desde el "Stata Journal" o del boletín técnico "Stata Technical Bulletin".
- Acceder a lugares de interés en el sitio Web de Stata.
- El sistema de ayuda para los comandos de Stata es una de las herramientas que más rápidamente puede familiarizar al usuario con el manejo de Stata. Alternativamente al sistema de ventanas, el usuario puede digitar en el cuadro de comandos help seguido del comando del cual desea información (Moschella & Rivas, 2009).

Por ejemplo al digitar en el cuadro de comandos: help describe emerge la siguiente Ventana.

Figura Nº 4



- Asimismo la opción de ayuda de Stata ofrece información sobre:
- La sintaxis completa y abreviada de letra(s) subrayadas) de cada comando,
- Descripción del comando solicitado
- Opciones adicionales para ejecutar el comando,
- Ejemplos sobre cómo usar el comando,
- Hipervínculos a otros comandos relacionados y/o similares y,
- El manual impreso de Stata en el que puede consultar los detalles sobre el comando.

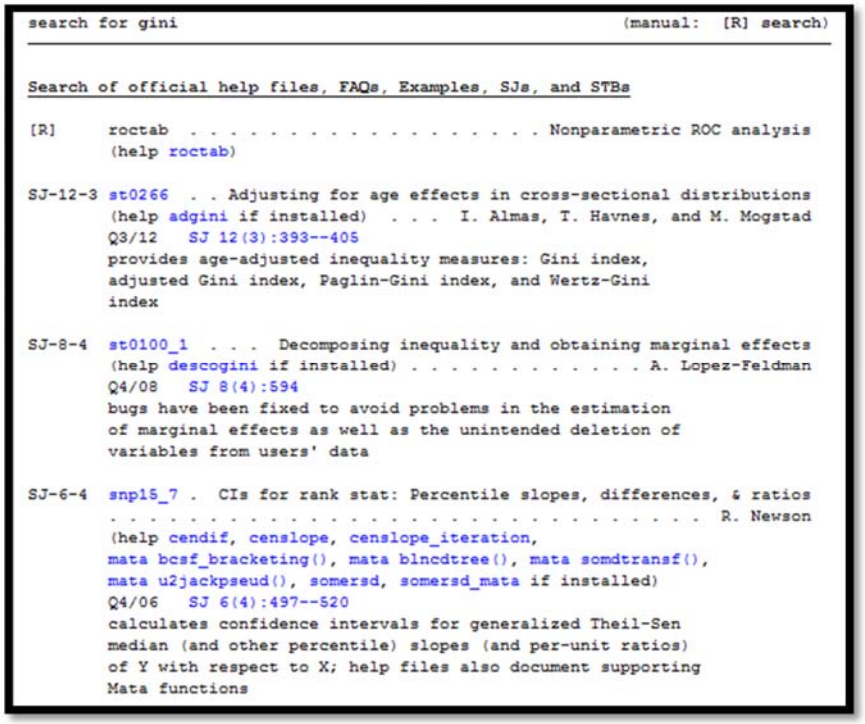
Con frecuencia, el usuario desconoce el nombre del comando específico que realiza algún procedimiento en Stata. En estos casos es conveniente realizar una búsqueda temática por medio del comando search. A través de este comando Stata realiza una búsqueda en línea en:

- Los ejemplos oficiales de Stata disponibles en su sitio web,
- El sitio de preguntas frecuentes "Frequently Asked Questions" de Stata,
- Ejemplos en línea compilados por la universidad de UCLA,
- Las referencias bibliográficas en "Stata Journal" y "Stata Technical Bulletin".

Por ejemplo, suponga que se quiere calcular en Stata el coeficiente de concentración gini (procedimiento muy conocido en economía y estadística), pero no se sabe si Stata realiza este cálculo y, además, si es posible hacerlo, no se conoce el comando para ejecutarlo. En estos casos el comando search (buscar) resulta de gran ayuda. Por ejemplo al escribir en el cuadro de comandos

search gini

Se despliega el siguiente cuadro de ayuda:



```
search for gini                                          (manual:  [R] search)

Search of official help files, FAQs, Examples, SJs, and STBs

[R]      roctab . . . . . . . . . . . . . . . . . Nonparametric ROC analysis
         (help roctab)

SJ-12-3  st0266 . . Adjusting for age effects in cross-sectional distributions
         (help adgini if installed) . . .  I. Almas, T. Havnes, and M. Mogstad
         Q3/12   SJ 12(3):393--405
         provides age-adjusted inequality measures: Gini index,
         adjusted Gini index, Paglin-Gini index, and Wertz-Gini
         index

SJ-8-4   st0100_1 . . . Decomposing inequality and obtaining marginal effects
         (help descogini if installed) . . . . . . . . . . . A. Lopez-Feldman
         Q4/08   SJ 8(4):594
         bugs have been fixed to avoid problems in the estimation
         of marginal effects as well as the unintended deletion of
         variables from users' data

SJ-6-4   snp15_7 . CIs for rank stat: Percentile slopes, differences, & ratios
         . . . . . . . . . . . . . . . . . . . . . . . . . . . . . R. Newson
         (help cendif, censlope, censlope_iteration,
         mata bcsf_bracketing(), mata blncdtree(), mata somdtransf(),
         mata u2jackpseud(), somersd, somersd_mata if installed)
         Q4/06   SJ 6(4):497--520
         calculates confidence intervals for generalized Theil-Sen
         median (and other percentile) slopes (and per-unit ratios)
         of Y with respect to X; help files also document supporting
         Mata functions
```

En el cuadro de ayuda aparecen en azul hipervínculos a sitios oficiales (Stata Journal "SJ", o Stata Technical Bulletin "STB") desde donde se pueden descargar macros relacionadas con el procedimiento que calcula el coeficiente de concentración Gini.

Automáticamente Stata hace actualizaciones periódicas del programa. Sin embargo el usuario puede pedir manualmente al programa que se actualice a través del comando update así:
update all
adoupdate, update

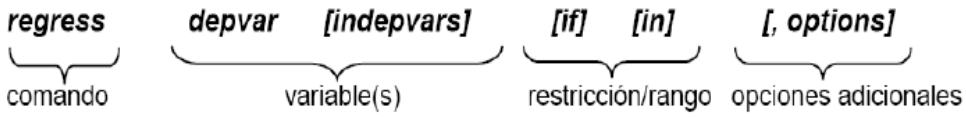
Antes de comenzar una sesión de trabajo es importante tener en cuenta que Stata opera a través de diferentes tipos de archivos.

III. ESTRUCTURA DE COMANDOS

La creación de las variables se realiza por medio del comando generate, los comandos en Stata no son necesarios escribirlos en su totalidad. La mayoría de los comandos pueden ser reducidos en un prefijo, para conocer el prefijo de cada comando escriba help nombre del comando y en la ayuda, aparecerá subrayado el nombre hasta cierto carácter indicando que puede usar solamente ese texto para ejecutar el comando, por ejemplo g es igual a generate.

[by varlist:] Command [varlist] [=exp] [if exp] [in range] [weight] [using filename] [, options]

Por ejemplo:



Cabe indicar que Stata distingue entre letras mayúsculas y minúsculas. Además todos los comandos del programa se deben escribir en letras minúsculas. De lo contrario el programa no lo reconoce. Los paréntesis cuadrados indican que no es un carácter obligatorio dependiendo el comando específico.
Es posible usar con Stata prefijos para algunos comandos, por ejemplo, el comando regress que permite realizar el procedimiento de regresión se puede ejecutar digitando solamente los tres primeros caracteres, es decir al tener reg ejecuta la misma función que al escribir regress. Para conocer mayor información sobre la estructura de los comandos de Stata, busque información así: help syntax

IV. LA BARRA DE HERRAMIENTAS DE STATA 14

Además, Stata presenta una barra de herramientas que permite realizar operaciones usuales como abrir un archivo, grabarlo, imprimir o ver alguna ventana en particular.

 Sirve para abrir una base de datos de Stata.

 Sirve para grabar en el disco la base de datos que está siendo usada.

 Imprime los gráficos o el contenido de la ventana Stata Viewer.

 Empieza una nueva bitácora, abre una existente, cierra o suspende la que se esté usando.

 Muestra una ventana Stata Viewer que esté oculta.

 Muestra la ventana Stata Results.

 Muestra el último gráfico creado.

 Abre un Do-File Editor o muestra la ventana de Do-File Editor que esté oculta (equivale a ctrl+8).

 Abre Stata Editor o muestra la ventana de Stata Editor que esté oculta (equivale a edit).

 Abre Stata Browser o muestra la ventana de Stata Browser que esté oculta (equivale a browse).

 Le dice a Stata que continúe la ejecución de un comando que ha sido detenido.

 Detiene Stata (equivale a la tecla q).

V. ASPECTOS INTRODUCTORIOS

A continuación, veremos en qué consiste una sesión de trabajo en Stata 14 y exploraremos algunos comandos. Al mismo tiempo conoceremos la forma interactiva de trabajo con Stata 14 utilizando la Barra de Menús y los GUI para realizar paralelamente algunas tareas.

Para abrir una base de datos desde el menú principal, seguimos la siguiente ruta: File/Open. En el cuadro de diálogo que aparece a continuación se elige el archivo deseado, que en este caso tiene la extensión de los archivos de datos nativos de Stata, dta. Como ejemplo de sesión abriremos el archivo **auto.dta**:

```
Contains data from C:\Users\FINANCE.BUSINESS\Dropbox\ECONOMETRIA I\LABORATORIO STATA\Sesion 1.rbr\auto.dta
  obs:            74                          1978 Automobile Data
 vars:            12                          14 Oct 2002 09:02
 size:         3,182                          (_dta has notes)
```

variable name	storage type	display format	value label	variable label
make	str18	%-18s		Make and Model
price	int	%8.0gc		Price
mpg	int	%8.0g		Mileage (mpg)
rep78	int	%8.0g		Repair Record 1978
headroom	float	%6.1f		Headroom (in.)
trunk	int	%8.0g		Trunk space (cu. ft.)
weight	int	%8.0gc		Weight (lbs.)
length	int	%8.0g		Length (in.)
turn	int	%8.0g		Turn Circle (ft.)
displacement	int	%8.0g		Displacement (cu. in.)
gear_ratio	float	%6.2f		Gear Ratio
foreign	byte	%8.0g	origin	Car type

```
Sorted by:  foreign
```

Este archivo contiene una base de datos de autos: 74 observaciones y 12 variables.

Observemos que se han suscitado cambios en las siguientes ventanas:
Review: use "C:\Users\FINANCE.BUSINESS\Dropbox\ECONOMETRIAI\LABORATORIO STATA\Sesion 1.rbr\auto.dta"
Variables: Aparece la lista con las variables del archivo auto.dta
Stata Results: use "C:\DATA\auto.dta", clear

Nota: Al ejecutar los comandos mediante estos menús además se registra en la ventana Stata Review el comando equivalente para la ventana Stata Command. Esta característica es muy útil cuando se aprende Stata porque es posible ejecutar un comando mediante la GUI y luego repetirlo empleando comandos.

Cuando cargamos en la memoria el archivo auto.dta mediante el menú Stata, éste ha incorporado en la ventana de resultados los comandos equivalentes que se hubiesen tenido que poner en Stata Command para obtener el mismo resultado. Los comandos se han almacenado en la ventana Review del mismo modo que sucedería si hubiésemos digitado los comandos en dicha ventana. Si bien trabajar con la barra de herramientas y con los menús desplegables es más intuitivo, para el

usuario experto es más rápido y sencillo potente trabajar directamente con los comandos (lo cual le permite emplear los archivos de ejecución y la programación avanzada)

Para ver una descripción rápida de los datos ingresamos describe. Para copiar a MSWord lo que acaba de aparecer en la ventana de resultados iluminamos dicho resultado y lo copiamos como texto o como tabla, luego de pegarlo le aplicamos formato indicando el tipo tamaño y fuente.

use "C:\Program Files\Stata9\auto"
describe

```
make            mpg             headroom        weight          turn            gear_ratio
price           rep78           trunk           length          displacement    foreign

. d

Contains data from auto.dta
  obs:            74                          1978 Automobile Data
 vars:            12                          13 Apr 2013 17:45
 size:         3,182                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
make            str18   %-18s                 Make and Model
price           int     %8.0gc                Price
mpg             int     %8.0g                 Mileage (mpg)
rep78           int     %8.0g                 Repair Record 1978
headroom        float   %6.1f                 Headroom (in.)
trunk           int     %8.0g                 Trunk space (cu. ft.)
weight          int     %8.0gc                Weight (lbs.)
length          int     %8.0g                 Length (in.)
turn            int     %8.0g                 Turn Circle (ft.)
displacement    int     %8.0g                 Displacement (cu. in.)
gear_ratio      float   %6.2f                 Gear Ratio
foreign         byte    %8.0g       origin    Car type
-------------------------------------------------------------------------------
Sorted by:  foreign
```

ds /*lista las variables en forma compacta*/ make

```
. ds
make            mpg             headroom        weight          turn            gear_ratio
price           rep78           trunk           length          displacement    foreign
```

codebook make mpg rep78 weight

codebook, escribe el contenido de las variables, indicando número de observaciones, valores perdidos, percentiles, entre otros. (Rojas, Brayan; Marcelo Gordillo, Darwin, 2013)

```
. codebook make mpg rep78 weight

-------------------------------------------------------------------------------
make                                                            Make and Model
-------------------------------------------------------------------------------

                  type:  string (str18), but longest is str17

          unique values:  74                       missing "":  0/74

               examples:  "Cad. Deville"
                          "Dodge Magnum"
                          "Merc. XR-7"
                          "Pont. Catalina"

                warning:  variable has embedded blanks

-------------------------------------------------------------------------------
mpg                                                             Mileage (mpg)
-------------------------------------------------------------------------------

                  type:  numeric (int)

                 range:  [12,41]                     units:  1
          unique values:  21                       missing .:  0/74

                  mean:   21.2973
             std. dev:    5.7855

            percentiles:        10%       25%       50%       75%       90%
                                 14        18        20        25        29
```

codebook foreign

```
. codebook foreign

-------------------------------------------------------------------------------
foreign                                                             Car type
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  origin

                 range:  [0,1]                       units:  1
          unique values:  2                      missing .:  0/74

            tabulation:  Freq.   Numeric  Label
                           52         0  Domestic
                           22         1  Foreign
```

```
-------------------------------------------------------------------------------
rep78                                                    Repair Record 1978
-------------------------------------------------------------------------------

                 type:  numeric (int)

                range:  [1,5]                        units:  1
        unique values:  5                    missing .:  5/74

           tabulation:  Freq.  Value
                           2  1
                           8  2
                          30  3
                          18  4
                          11  5
                           5  .

-------------------------------------------------------------------------------
weight                                                   Weight (lbs.)
-------------------------------------------------------------------------------

                 type:  numeric (int)

                range:  [1760,4840]                  units:  10
        unique values:  64                   missing .:  0/74

                 mean:   3019.46
            std. dev:   777.194

          percentiles:        10%       25%       50%       75%       90%
                             2020      2240      3190      3600      4060
```

.

A continuación se muestra una parte de nuestra data:

list make mpg in 1/10

```
. list make mpg in 1/10

     +-------------------+
     |  make         mpg |
     |-------------------|
  1. | AMC Concord    22 |
  2. | AMC Pacer      17 |
  3. | AMC Spirit     22 |
  4. | Buick Century  20 |
  5. | Buick Electra  15 |
     |-------------------|
  6. | Buick LeSabre  18 |
  7. | Buick Opel     26 |
  8. | Buick Regal    20 |
  9. | Buick Riviera  16 |
 10. | Buick Skylark  19 |
     +-------------------+
```

¿Qué carros tienen el menor millaje por galón?

Para realizar ello se utiliza el comando sort para ordenarlos de menor a mayor según mpg.

sort mpg

list make mpg in 1/10

	make	mpg
1.	Linc. Continental	12
2.	Linc. Mark V	12
3.	Cad. Deville	14
4.	Cad. Eldorado	14
5.	Peugeot 604	14
6.	Merc. Cougar	14
7.	Linc. Versailles	14
8.	Merc. XR-7	14
9.	Buick Electra	15
10.	Merc. Marquis	15

¿Cuáles son los cinco autos con mayor millaje por galón?

list make mpg in -10/-1

	make	mpg
65.	Ford Fiesta	28
66.	Plym. Arrow	28
67.	Chev. Chevette	29
68.	Mazda GLC	30
69.	Dodge Colt	30
70.	Toyota Corolla	31
71.	Plym. Champ	34
72.	Subaru	35
73.	Datsun 210	35
74.	VW Diesel	41

Para ver los datos tal como si los viésemos en MSExcel digitamos browse y aparece la ventana Stata Editor.

Browse 

Para editar los datos escribimos edit o pulsamos el botón correspondiente[2]:

Edit 

VI. TIPOS Y FORMATOS DE VARIABLES

Una de las preguntas comunes en el manejo de un software estadístico es cómo el programa clasifica o categoriza las variables, es decir que formato es posible asignarle a una variable, para ello es necesario primero que el usuario tenga claro el tipo de variable. Las variables se pueden dividir de acuerdo al siguiente esquema (Rojas, Brayan; Marcelo Gordillo, Darwin, 2013):

float	números reales en formato 8,5 (8 cifras enteras, cinco decimales)
double	números reales en formato 16,5
byte	enteros entre –127 y 100
int	enteros entre –32767 y 32740
long	enteros entre –3147483647 y 2147483620

Stata por defecto le asigna formato float a una variable de datos nueva. Intercooled Stata14.0 soporta cadenas de hasta 80 caracteres de largo.

str1 cadenas de 1 carácter
str80 cadenas de 80 caracteres

[2] Nota: Cuando las ventanas Browser y Edit se encuentran abiertas es imposible ingresar comandos, puesto que la barra Stata Command desaparece.

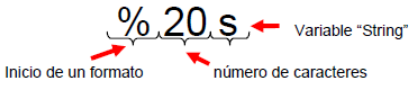
El número que aparece después del símbolo % es el número máximo de dígitos enteros o ancho que soporta el formato y el número a la derecha índica el número de decimales, posteriormente se encuentra una letra. Donde [f] es aproximación al entero más cercano, [e] indica notación científica y [g] indica decimales.

Stata por defecto selecciona el formato FLOAT, el otro tipo de variables son las variables alfanuméricas, estas variables en las que se encuentran principalmente las variables cualitativas, Stata define un formato especial para ellas, y es el formato STRING, %str# es la visualización de este formato, en el cual el carácter # indica el largo de la cadena.

FORMATO DE LAS VARIABLES

El formato de las variables hace referencias a la forma como son almacenadas y desplegadas las variables en STATA. Para cambiar el formato de una de una variable a través del lenguaje de sintaxis debe tener en cuenta que el formato de toda variable siempre antecedido por el símbolo "%".

Variables de cadena



Variable numérica

Si desea cambiar el formato de una variable utilice el comando recast.

sysuse auto
describe Price
recast float price

Para mayor información: help data_types y help recast

V. ESTADÍSTICAS DESCRIPTIVAS

¿Cómo no estoy familiarizado con los precios de 1978, cuál es el precio promedio de los carros en esta base de datos?

summarize price

```
. summarize price
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	74	6165.257	2949.496	3291	15906

summarize funciona como list, pero sin argumentos, nos da un resumen de toda la data:

summarize

```
. help summarize

. summarize price
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	74	6165.257	2949.496	3291	15906

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

```
.
```

Observación: la variable make tiene 0 observaciones debido a que es una variable de cadena (string), calcular una media bajo este comando está indefinido pero no

es un error hacerlo. La variable rep78 sólo tiene 69 observaciones porque no tiene registro para 5 carros.

¿Cuál es el precio promedio de los carros que se encuentran por encima y por debajo de la media de mpg?

summarize price if mpg<21.3

```
. summarize price if mpg<21.3
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	43	7091.86	3425.019	3291	15906

```
.
```

summarize price if mpg>=21.3

```
summarize price if mpg>=21.3
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	31	4879.968	1344.659	3299	9735

if puede utilizarse como sufijo para casi todos los comandos. Esta es una de las características más útiles en Stata.

¿Cuál es la mediana de mpg?

summarize mpg, detail

```
. summarize mpg, detail

                          Mileage (mpg)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           12            12
 5%           14            12
10%           14            14       Obs                  74
25%           18            14       Sum of Wgt.          74

50%           20                     Mean            21.2973
                      Largest        Std. Dev.       5.785503
75%           25            34
90%           29            35       Variance        33.47205
95%           34            35       Skewness        .9487176
99%           41            41       Kurtosis        3.975005

.
```

Nuestra base de datos contiene la variable foreign que esta codificada de la siguiente manera: 0 si el carro ha sido fabricado los Estados Unidos o Canada, y 1 si el carros ha sido fabricado en otra parte. ¿Existen diferencias de precio o millaje explicadas por el origen de fabricación? Para saberlo tenemos que estimar las estadísticas de resumen para las variables price y MPG en los dos casos se recoge la variable foreign.

Existen dos soluciones para este problema:

summarize price mpg if foreign==0
summarize price mpg if foreign==1
O, introduciendo las siguientes líneas de comandos:
sort foreign
by foreign: summarize price mpg

```
. by foreign: summarize price mpg

-------------------------------------------------------------------------------
-> foreign = Nacional

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         52    6072.423    3097.104       3291      15906
         mpg |         52    19.82692    4.743297         12         34

-------------------------------------------------------------------------------
-> foreign = Extanjero

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         22    6384.682    2621.915       3748      12990
         mpg |         22    24.77273    6.611187         14         41
```

Para descartar si el promedio de los carros "domésticos" extranjeros es diferente. Lo que queremos ahora es saber si son "estadísticamente" diferente. Para ello haremos un contraste de hipótesis para verificar si las medias de ambos grupos son iguales.

ttest mpg, by(foreign)

```
Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      52    19.82692     .657777    4.743297    18.50638    21.14747
 Foreign |      22    24.77273     1.40951    6.611187    21.84149    27.70396
---------+--------------------------------------------------------------------
combined |      74      21.2973    .6725511    5.785503     19.9569    22.63769
---------+--------------------------------------------------------------------
    diff |           -4.945804    1.362162               -7.661225   -2.230384
------------------------------------------------------------------------------
    diff = mean(Domestic) - mean(Foreign)                         t =  -3.6308
Ho: diff = 0                                     degrees of freedom =       72

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0003         Pr(|T| > |t|) = 0.0005          Pr(T > t) = 0.9997
```

A partir de esta prueba hemos establecido que los carros domésticos en 1978 tenían un menor millaje de gasolina que los carros extranjeros. Ahora, lo que queremos es saber el número de carros domésticos y extranjeros.

tabulate foreign

```
. tabulate foreign
```

Car type	Freq.	Percent	Cum.
Domestic	52	70.27	70.27
Foreign	22	29.73	100.00
Total	74	100.00	

La base de datos contiene la variable rep78 que a registrar la frecuencia de mantenimiento para cada caso (1 = mala,..., 5 = excelente). ¿Cómo ha sido el mantenimiento de los carros de la muestra?

tabulate rep78

```
. tabulate rep78
```

Repair Record 1978	Freq.	Percent	Cum.
1	2	2.90	2.90
2	8	11.59	14.49
3	30	43.48	57.97
4	18	26.09	84.06
5	11	15.94	100.00
Total	69	100.00	

Tenemos 74 carros, sólo 69 tienen registrada la variable rep78 . Queremos conocer los carros para los cuales esta información no existe[3].

list make if rep78>=.

	make
3.	AMC Spirit
7.	Buick Opel
45.	Plym. Sapporo
51.	Pont. Phoenix
64.	Peugeot 604

Queremos saber si existen también diferencias entre las frecuencia de mantenimiento explicadas por el origen de fabricación. Comparemos los registros

[3] Nota: list make if rep78>=. Es equivalente a **list make if missing(rep78)**

para los carros domésticos y extranjeros, es decir, hagamos una tabla en la que comparemos dos variables rep78 y foreign.

tabulate rep78 foreign

```
. tabulate rep78 foreign
```

Repair Record 1978	Car type Domestic	Foreign	Total
1	2	0	2
2	8	0	8
3	27	3	30
4	9	9	18
5	2	9	11
Total	48	21	69

Parece que los carros domésticos tienen una menor frecuencia de mantenimiento. A continuación, queremos determinar si esta diferencia es estadísticamente significativa. Para ello realizaremos un test chi2 (a pesar de que no se cumple la condición de que debe haber como mínimo cinco observaciones en cada celda de la tabla).

tabulate rep78 foreign, chi2

Hemos encontrado que la frecuencia de mantenimiento es diferente de por el origen de fabricación. Podemos inferir que los carros domésticos tenían un menor nivel de mantenimiento en 1978 (Moschella & Rivas, 2009).

Repair Record 1978	Car type Domestic	Foreign	Total
1	2	0	2
2	8	0	8
3	27	3	30
4	9	9	18
5	2	9	11
Total	48	21	69

Pearson chi2(4) = 27.2640 Pr = 0.000

Matrices de correlación
¿Cuál es la correlación entre MPG y el peso de un carro?

correlate mpg weight

```
. correlate mpg weight
(obs=74)

             |      mpg   weight
-------------+------------------
         mpg |   1.0000
      weight |  -0.8072   1.0000
```

Comparemos esta correlación para los carros domésticos y extranjeros:

```
. correlate mpg weight if foreign==0
(obs=52)

             |      mpg   weight
-------------+------------------
         mpg |   1.0000
      weight |  -0.8759   1.0000
```

correlate mpg weight if foreign==1

```
. correlate mpg weight if foreign==1
(obs=22)

             |      mpg   weight
-------------+------------------
         mpg |   1.0000
      weight |  -0.6829   1.0000
```

También pudimos obtener el mismo resultado tipeando

by foreign:correlate mpg weight

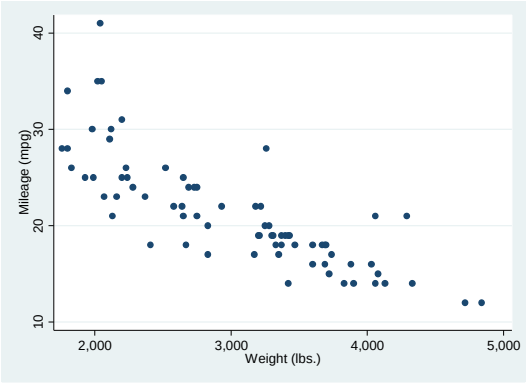
Podemos estimar matrices de correlación con tantas variables como queramos.

```
-> foreign = Domestic
(obs=52)

             |      mpg   weight
-------------+------------------
         mpg |   1.0000
      weight |  -0.8759   1.0000


-> foreign = Foreign
(obs=22)

             |      mpg   weight
-------------+------------------
         mpg |   1.0000
      weight |  -0.6829   1.0000
```

VI. GENERACIÓN BÁSICA DE GRÁFICOS

Sabemos que el millaje promedio entre los carros domésticos y extranjeros es diferente. Hemos visto también que el origen de fabricación explica otras diferencias, tales como la frecuencia de reparación. Por otro lado, hemos encontrado una correlación negativa entre MPG y el peso del carro (como era de esperarse) pero esta correlación parece ser más fuerte cuando analizamos carros domésticos. A continuación examinaremos, con la intención de más adelante modelar, la relación entre MPG y el peso. Comenzaremos graficando un ploteo simple (Moschella & Rivas, 2009)

Para generar gráficos manualmente, tenemos que situarnos en el menú principal: ***Graphics/Easy graphs/Scatter plot***, y en el cuadro de diálogo, con el cursor en la casilla X variable pulsamos mpg y en la casilla Y variable pulsamos weight. Luego de pulsar el botón OK, nos mostrará la siguiente pantalla (Moschella & Rivas, 2009).



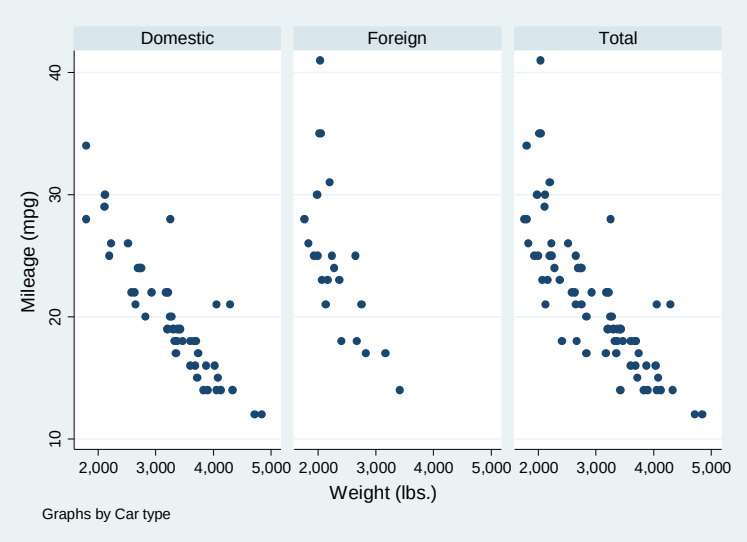
Tambíen podemos tipear la siguiente orden:

scatter mpg weight

scatter y x dibuja un gráfico de y contra x. Notamos que la relación a primera vista no es lineal.

A continuación, realizaremos dos gráficos separados para los carros domésticos y extranjeros.

sort foreign
scatter mpg weight, by(foreign, total row(1))



La relación no es únicamente no lineal, sino que también parece diferir para los carros domésticos y extranjeros.

VIII. COMBINACIÓN DE BASES DE DATOS

La combinación de bases de datos es un problema muy común para el investigador o el analista de información, Stata le permite realizar diferentes tipos de fusiones de bases de datos, a continuación se presentan los dos formatos más importantes, la adición vertical (merge) y horizontal (append). Suponga que usted debe consolidar la información de las siguientes bases de datos para generar un reporte. Como parte del trabajo del investigador es acopiar y consolidar la información, este problema es muy común. Para combinar datos se emplean los comandos append, merge y joinby (Rojas & Gordillo, 2012).

BASE 1

Nombre	Sexo	Econometria	Gerencia
Alfredo	1	15	14
Juan	1	14	17
Maria	0	16	18
Josefina	0	12	12
Geraldine	0	18	11
Pedro	1	15	19
Alberto	1	18	13
Carlos	1	13	12

BASE 3

Nombre	Sexo	Econometria	Finanzas
Alfredo	1	15	14
Juan	1	12	12
Maria	0	18	11
Josefina	0	15	19
Geraldine	0	18	13
Pedro	1	13	12
Alberto	1	14	17
Carlos	1	16	18

BASE 2

Nombre	Sexo	Econometria	Gerencia
Luis	1	15	14
Renzo	1	14	17
Mariella	0	16	18
Carla	0	12	12
Petronila	0	18	11
Daniel	1	15	19
Rafael	1	18	13
Sandro	1	13	12

BASE 4

Nombre	Sexo	Finanzas
Luis	1	15
Renzo	1	14
Mariella	0	16
Carla	0	12
Petronila	0	18
Daniel	1	15
Rafael	1	18
Sandro	1	13

El comando **append y merge** nos ayudara a unir bases de datos integrándolas en una sola. append, pegara hacia abajo o verticalmente y Merge, pegara hacia el costado o de forma horizontal.

```
clear
use base1, clear
list
use base2, clear
list
use base3, clear
list
```

use base4, clear
list

Vamos a empezar observando cada una de las bases de datos que tenemos

```
. use base2, clear

. list
```

	Nombre	Sexo	Econom~a	Gerencia
1.	Alfredo	1	15	14
2.	Juan	1	14	17
3.	Maria	0	16	18
4.	Josefina	0	12	12
5.	Geraldine	0	18	11
6.	Pedro	1	15	19
7.	Alberto	1	18	13
8.	Carlos	1	13	12

```
. use base3, clear

. list
```

	Nombre	Sexo	Econom~a	Finanzas
1.	Alfredo	1	15	14
2.	Juan	1	12	12
3.	Maria	0	18	11
4.	Josefina	0	15	19
5.	Geraldine	0	18	13
6.	Pedro	1	13	12
7.	Alberto	1	14	17
8.	Carlos	1	16	18

```
. use base4, clear

. list
```

	Nombre	Sexo	Finanzas
1.	Luis	1	15
2.	Renzo	1	14
3.	Mariella	0	16
4.	Carla	0	12
5.	Petronila	0	18
6.	Daniel	1	15
7.	Rafael	1	18
8.	Sandro	1	13

La base de datos Base1 tiene los mismos campos (columnas) que la base de datos Base2, pero diferentes filas, sería útil, unir ambas bases. Abramos entonces, la base de datos Base1 y peguémosla con la base de datos Base2, una unión vertical.

Hagamos lo mismo con las bases de datos Base3 y Base4 y observemos los resultados:

	Nombre	Sexo	Econom~a	Finanzas
1.	Alfredo	1	15	14
2.	Juan	1	12	12
3.	Maria	0	18	11
4.	Josefina	0	15	19
5.	Geraldine	0	18	13
6.	Pedro	1	13	12
7.	Alberto	1	14	17
8.	Carlos	1	16	18
9.	Luis	1	.	15
10.	Renzo	1	.	14
11.	Mariella	0	.	16
12.	Carla	0	.	12
13.	Petronila	0	.	18
14.	Daniel	1	.	15
15.	Rafael	1	.	18
16.	Sandro	1	.	13

Ahora si resulto bien la unión vertical. Veamos la base12 que teníamos antes. Ahora mi interés es fusionar ambas bases de datos, para ello, primero debemos ordenar ambas bases según la variable con la que vamos a fusionar (la variable común).

```
save base34s.dta, replace
use base12.dta, clear
list
sort nombre
list
save base12s.dta, replace
```

Ya tenemos las 2 bases de datos ordenadas, ahora vamos a fusionarlas

```
clear
use base12s.dta, clear
list
merge Nombre using base34s.dta
list
save basetotal.dta, replace
```

```
. list
```

	Nombre	Sexo	Econom~a	Gerencia	Finanzas	_merge
1.	Alberto	1	18	13	17	3
2.	Alfredo	1	15	14	14	3
3.	Carla	0	12	12	12	3
4.	Carlos	1	13	12	18	3
5.	Daniel	1	15	19	15	3
6.	Geraldine	0	18	11	13	3
7.	Josefina	0	12	12	19	3
8.	Juan	1	14	17	12	3
9.	Luis	1	15	14	15	3
10.	Maria	0	16	18	11	3
11.	Mariella	0	16	18	16	3
12.	Pedro	1	15	19	12	3
13.	Petronila	0	18	11	18	3
14.	Rafael	1	18	13	18	3
15.	Renzo	1	14	17	14	3
16.	Sandro	1	13	12	13	3

CREAR UNA BASE DE DATOS A PARTIR DE OTRA

Para esta rutina se utiliza el comando **collapse** que toma una base de datos original y crea una nueva que contiene estadísticas de resumen de la base de datos original. Asimismo agrega etiquetas de las variables en esta nueva base de datos. Debido a que el diagrama de sintaxis para el collapse hace que su uso parezca más complicado de lo que es, sin embargo collapse se explica mejor con ejemplos (STATA14, 2015).

Usando datos acerca del rendimiento de los alumnos GPA de un colegio.

use http://www.stata-press.com/data/r13/college

```
Contains data from C:\Users\formulacion1\Downloads\college.dta
  obs:            12
 vars:             4                          3 Jan 2011 12:05
 size:           120
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
gpa             float   %9.0g                 gpa for this year
hour            int     %9.0g                 Total academic hours
year            int     %9.0g                 1 = freshman, 2 = sophomore, 3 = junior, 4 =
                                                senior
number          int     %9.0g                 number of students
-------------------------------------------------------------------------------
Sorted by:  year
```

list, sep(4)

A continuación vemos la base de datos que está ordenada según años. Además nos muestra doce observaciones por variable.

	gpa	hour	year	number
1.	3.2	30	1	3
2.	3.5	34	1	2
3.	2.8	28	1	9
4.	2.1	30	1	4
5.	3.8	29	2	3
6.	2.5	30	2	4
7.	2.9	35	2	5
8.	3.7	30	3	4
9.	2.2	35	3	2
10.	3.3	33	3	3
11.	3.4	32	4	5
12.	2.9	31	4	2

Para obtener una base de datos que contiene el percentil 25 de GPA para cada año, escribimos

collapse (p25) gpa [fw=number], by(year)

Utilizamos pesos de frecuencia. Ahora vamos a crear un conjunto de datos que contiene la media de GPA y hora para cada año. Nosotros no tiene que escribir (media) para especificar que queremos que la media debido a que el medio se informó de forma predeterminada .

use http://www.stata-press.com/data/r13/college, clear

Tenemos los datos del censo que contienen información sobre la edad mediana de cada estado , la tasa de matrimonio y el divorcio tarifa. Queremos formar un nuevo conjunto de datos que contiene diversas estadísticas de resumen, por regiones, de las variables :

use http://www.stata-press.com/data/r13/census5, clear

describe

```
Contains data from C:\Users\formulacion1\Downloads\census5.dta
  obs:            50                          1980 Census data by state
 vars:             7                          6 Apr 2011 15:43
 size:         1,700
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
state           str14   %14s                  State
state2          str2    %-2s                  Two-letter state abbreviation
region          int     %8.0g      cenreg     Census region
pop             long    %10.0g                Population
median_age      float   %9.2f                 Median age
marriage_rate   float   %9.0g
divorce_rate    float   %9.0g
-------------------------------------------------------------------------------
Sorted by:  region
```

Nos dice además que la base de datos esta ordenada por región.

collapse (median) median_age marriage divorce (mean) avgmrate=marriage > avgdrate=divorce [aw=pop], by(region)

Con esta secuencia de rutinas, le estamos diciendo al programa, que cree una nueva base de datos que contenga la media y la mediana. Se está cambiando de nombre de las variables marriage por avgmrate divorce por avgdrate.

list

	region	median~e	marria~e	divorc~e	avgmrate	avgdrate
1.	NE	31.90	.0080657	.0035295	.0081472	.0035359
2.	N Cntrl	29.90	.0093821	.0048636	.0096701	.004961
3.	South	29.60	.0112609	.0065792	.0117082	.0059439
4.	West	29.90	.0089093	.0056423	.0125199	.0063464

describe

```
Contains data
  obs:             4                          1980 Census data by state
 vars:             6
 size:            88
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
region          int     %8.0g      cenreg     Census region
median_age      float   %9.2f                 (p 50) median_age
marriage_rate   float   %9.0g                 (p 50) marriage_rate
divorce_rate    float   %9.0g                 (p 50) divorce_rate
avgmrate        float   %9.0g                 (mean) marriage_rate
avgdrate        float   %9.0g                 (mean) divorce_rate
-------------------------------------------------------------------------------
Sorted by:  region
     Note:  dataset has changed since last saved
```

ANEXO

Stata se instala por defecto en C:\STATA\ y guarda los datos con los que se trabaje en C:\DATA\ salvo que dicha configuración de archivos haya sido cambiada. Para conocer con que directorio está trabajando Stata se utiliza el comando pwd

pwd
E:\Stata14

Con el comando sysdir se pueden visualizar los directorios que STATA emplea para guardar el programa y la información.

```
. sysdir
   STATA:  D:\Stata1332\
    BASE:  D:\Stata1332\ado\base\
    SITE:  D:\Stata1332\ado\site\
    PLUS:  c:\ado\plus\
PERSONAL:  c:\ado\personal\
OLDPLACE:  c:\ado\

.
```

Con el comando cd podemos cambiar de directorio donde se guardan los datos, aunque estos cambios solo serán válidos por la sesión en uso.

cd C:\
C:\
pwd
C:\

- Para volver a fijar como directorio en uso el directorio anterior: cd C:\\Cursos\Stata\Sesion02
- Para crear un nuevo directorio se emplea el comando mkdir:
- Para obtener un listado de todos los archivos del directorio en uso del comado dir
- Para restringir el listado a los datos: dir *.dta

- Es posible también copiar los datos de un archivo: copy auto.dta auto.bak
- También es posible borrar archivos: erase auto.bak

GUARDANDO BITÁCORAS

STATA permite guardar un registro de los comandos y los resultados. Para crear un archivo de bitácora se usa el comando log using:

log using bitacora

Por defecto se guarda en el directorio en uso y en formato SMCL (que es el que usa STATA) para presentar los resultados.

use "C:\cursos\Stata8\Sesion01\auto.dta", clear describe

Para parar momentáneamente el registro de la bitácora: log off

Este comando no será registrado: summ

Para reanudar el registro de la bitácora: log on

Este comando si será registrado: tab rep78

Para detener la bitácora: log close

Para reanudar una bitácora: log using bitácora, append

Para sobrescribir una bitácora: log using bitacora, replace

Para ver una bitácora: type bitacora.smcl

Si lo único que se desea es guardar los comandos debe emplearse cmdlog. Esta opción es especialmente útil cuando lo que se busca es crear un archivo do.

cmdlog using C:\cursos\Stata14\Sesion01\comusados help log

USO DEL MENU AYUDA DE STATA.

Comando help; Este es uno de los comandos más importantes de Stata, pues presenta la sintaxis de los comandos así como ejemplos de cómo se usan. Para pedirle ayuda a Stata sobre un comando se escribe en la ventana de comandos help (o hel o he) seguido del nombre del comando que queramos conocer (Moschella & Rivas, 2009).

Abramos el archivo auto.dta y luego escribamos:

help summarize

Observemos que el comando y sus opciones están en letra celeste mientras que la mayor parte del resto de la sintaxis va en letra negra[4]. En el nombre del comando se observa que las dos primeras letras están subrayadas, lo que indica que podemos en lugar de escribir el comando como summarize podemos escribir su (o cualquier forma intermedia).

Notemos que todo lo que va entre corchetes es optativo y que en la sintaxis de los comandos primero se ponen las variables, luego el peso, los condicionales (if) preceden a los rangos (in) y les siguen las opciones después de un coma:

[varlist]	Es la lista de variables.
[weight]	Son los pesos o ponderaciones.
[if exp]	Permite seleccionar la muestra donde exp es una expresión lógica.
[in range]	Permite seleccionar la muestra donde range es un rango de los datos.

La sintaxis de los comandos Stata tienen un formato común:

[by lista de var:] comando lista de var [if expresion] [in rango][ponderadores] [using nombre del archivo], [opciones]

Sin embargo es frecuente usar la versión mucho más simple como:

[by lista de var:] comandolista de var [if expresion], [opciones]

El prefijo by permite aplicar el mismo comando separando la base de datos en subgrupos de nidos por lista de variables. Posteriormente viene el comando seguido por una segunda lista de var a las cuales se les aplicara el comando elegido.

Los datos utilizados para evaluar el comando pueden ser limitados con las opciones if e in. Las opciones específicas al comando tienen que ser precedidas por una coma (Moschella & Rivas, 2009).

Otra información clave es la forma en que podemos obtener ayuda. Todos los comandos Stata tienen información acerca de la manera en que deben utilizarse (sintaxis y opciones); para acceder a ella es solo cuestión de escribir la palabra help

[4] La letra celeste se reserva para los hipervínculos.

seguida por el nombre del comando en la ventana de comandos de Stata. Si no conoce el nombre del comando que realiza la tarea que tiene en mente, escriba la palabra findit seguida por una palabra que este relacionada con dicha tarea. Este comando busca en toda la documentación tanto interna como aquella que se encuentra en la página red de Stata.

Stata se actualiza casi continuamente, los usuarios pueden escribir programas y mandarlos al archivo de SSC (Software Components), por lo tanto es necesario hacer actualizaciones de forma regular. El comando update query le indicará si es necesario hacer actualizaciones (Moschella & Rivas, 2009).

DESCRIPCION DE COMANDOS

Bibliografía

Moschella, M., & Rivas, J. (2009). Stata 9.0 para economistas. *Notas de Clase del Curso Stata para Economistas*. Lima: INFOPUCP.

Rojas, B., & Gordillo, D. (2012). *Introducción a la modelación de Base de Datos con Stata.* Santiago: Software Shop .

Rojas, Brayan; Marcelo Gordillo, Darwin. (2013). *Introducción al Análisis y Modelación de Datos Con Stata 12.* SOFTWARE shop.

STATA14, S. u. (2015). *Stata: Release 14. Statistical Software.* Stata Corp. Obtenido de http://www.stata.com/manuals14/u.pdf